Description

DOCUMENT RETRIEVAL SYSTEM WITH ACCESS CONTROL

TECHNICAL FIELD

The invention relates to electronic document retrieval and in particular to access control for documents available on storage devices located remotely from each other.

BACKGROUND ART

In communication networks, document servers, i.e. electronic document storage devices such as large disk drives, are frequently located remotely from each other. In large companies, with plants and offices scattered in many different cities, a computer network is often designed to link all or most locations. The network frequently includes a search engine or query server having an index of every word in every document which is selected for electronic availability, together with indexes for every word of such documents, and with pointers identifying the full document and its server where it may be accessed by an address known as a URL. Users with terminals on the network can address the query server with questions phrased in terms of key words and obtain documents which contain the key words. The questions are usually phrased or interpreted by the query server with query operators. The index at the query server is consulted to determine if the keywords are in the index, how many times they appear, and the number of documents which are responsive to the question, as interpreted by the search engine at the query server. However, a user is not given access to those documents which are beyond his or her access level.

In the prior art, the query server contained one list having the access level of each user. The index at the query server contained the access level associated with each corresponding document. Access was provided only to those documents for which the access level of the

user was appropriate by matching the two lists. The problem here was that the query server had to associate a security level with each document in the index, a cumbersome storage task. In the simplest case, a user is either given permission to search the database, or access is denied, with no middle ground.

Variations of the above approach exist, but in most instances there is a comparison of two lists - the user list, with associated access levels, and the document list, with associated access levels. The document list contains the access level for each appearance of each document in the index. An object of the invention was to devise an access control system which enables secure searching without having to store any access information in the database associated with the search engine.

A further object of the invention is to allow changes in a document server's access control list to be immediately reflected in searches of documents within that document server.

A still further object of the invention is to allow a single centralized index of multiple document servers to be created, whereby searches of this central collection will only return titles of documents that a user has access to, with access control being determined at the remote document servers which contain relevant documents.

SUMMARY OF THE INVENTION

The above object is achieved with a document retrieval system, with access control, in which the documents are stored in a distributed manner over a plurality of servers in a network, termed "web servers", but no access levels are associated with the documents or with the index at a query server. Instead of multiple control lists, a user enters, either manually or automatically, his or her user identification, together with the query to be searched. The search engine at the

query server receives the question and interprets the query operators to determine the number of hits responsive to the question. Each hit is associated with a document, in electronic form, located at a particular server by means of a pointer, known as a URL. However, before the hits are returned to the user, the hits are "screened" by determining from the web server whether the user has access using an access control list associated with the web server. The list associates user identification with URLs to which the user may have access.

The search engine will not report the presence of the documents for which the access level is insufficient. The web server returns documents for which the access level is compatible. Hence, the net result is that the user appears only to be able to search documents that the user has access to.

An advantage of the present invention is that the security of each document is always consistent between the web server and the search index.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a plan view of the document retrieval and access control system of the present invention.

Fig. 2 is a plan view of a first embodiment of an access control system in accord with the present invention.

Fig. 3 is a plan view of a second embodiment of an access control system in accord with the present invention.

Fig. 4 is a plan view of a third embodiment of an access control system in accord with the present invention.

BEST MODE FOR CARRYING OUT THE INVENTION

With reference to Fig. 1, a communications network, such as the Internet or a corporate intranet is indicated by the data bus line 11, a high speed conduit for digital data. Part of the network includes a query server 13 which is seen to comprise a search engine 15 which is connected to the text index database 17. The search engine is a high speed processor connected to the network by communications link 11. The search engine has the task of going to each document address in the network, known as a URL, combing through each document associated with the URL and indexing such words in a text index 17. A "URL" is an address or pointer to a document, or a file, or a record in a database, in other words to a piece of information which has been stored at a site known as a web server 23. The initials URL are an abbreviation for "uniform resource locator", recognized by Internet and intranet servers.

The URL is a string of ASCII characters with three common parts, a protocol indicator, a host server name, and a directory and file name, assuming that a file is the search target. An example would be http://federalexec.justice.gov/fbi/agents/cellular/pagers.html. In the example, the protocol is "http" which is hypertext transfer protocol, a common protocol which allows linking of files. The host server is "federalexec.justice.gov/fbi/. The document and its directory is "fbi/agents/cellular/pagers.html. The present invention takes advantage of the hierarchical structure of URLs by allowing access to all items of information specified in the initial portions of the URL for each user, i.e. a partial URL. The more detail specified in a URL, the lower the level of access. For example, the access level associated with http://federalexec.justice.gov/fbi/ might allow access to all files and data in the fbi server, while the further specification of "/agents" would further specify a level

of access. In the present invention, a list of users would have each user associated with URLs, or partial URLs which that user could access. The http protocol is particularly useful because it works with "browsers", i.e. software programs which allow the formatting of documents in a uniform manner which can be read by any computer or terminal which can run the browser software. Thus, a terminal or computer need not have access to the word processing program in which the document was prepared. Perhaps the best known feature of http protocol is the "hyperlink" feature, allowing a user to jump from a word or symbol in one document to another URL which expands upon the word or symbol. Another type of protocol which is commonly used is "ftp" or "fle transfer protocol". This allows direct access to computer files on designated servers and is not necessarily oriented to documents with hyperlinks, like http protocol.

A text index at a query server, lists the words found in documents accessible to the server. In response to a search request, the query server interprets the request and produces the number of hits for the search terms together with the associated URLs for the information. Thus, the query server holds information on all documents of all Internet/intranet sites and can produce corresponding URLs after a search. However, a user may not have proper access level for all of the documents found. In accord with the present invention, the user sees only those documents for which he has proper access.

A typical web site 21 includes a web server 23 and a document storage device 25. The web server 23 is a high speed processor and the storage device 25 is a disk drive. An access control list server 27 may be stored on storage device 25 or may have its own auxiliary storage device, as indicated in Fig. 1. If a separate storage device is used, such as a disk drive, it is also controlled by the web server 23. The web site 21 communicates with other web sites, not shown, which are also

on the network and joined by one or more communication links, such as data bus line 11. The storage device 25 holds electronic versions of documents which are available for searching and retrieval, but without any access control information.

In a corporate environment, the web site 21 may hold documents from the single plant or factory of a corporation. Other plants and factories have similar web sites which are all linked in a network known as an intranet. Access to documents is limited to persons who have proper authorization. Such authorization is maintained in the access control list server 27 associated with each web site. The list server 27 contains user IDs and the list of URLs or partial URLs that each user may access. In another example, corporate payroll record documents might be accessible to all department level managers and their supervisors, plus all members of the payroll and accounting departments. All other corporate employees would not have access to payroll records and so would be excluded from payroll documents available on storage device 25.

In operation, a user would send a query to search engine 15 which would interpret the query. An optional communications link 31 is provided to the access control list server 27 to determine whether the user may access web site 21 which has certain corporate documents in the search area under request. Assuming the user has initial access to the home page of web site 21 the search progresses by applying search terms to the index on storage device 17 which has pointers to text documents, such as URLs, found in the storage device 25 within web site 21. Assuming that payroll information is being requested and assuming that the payroll information is stored on storage device 25 which is accessed through the web server 23, the user identification is passed along to the web server 23. The web server 23 has access to the access control list server 27. The text index 17 has

identified documents in storage device 25. The access control list server 27 prevents the web server 23 from delivering any documents where the user identification indicates that the user does not authorization. Only those documents are pulled up for which the user has authorization. Those documents are then reported by the web server 23 to the search engine 15 which, in turn, reports the titles or bibliographic abstracts to the user. It should be noted that the user does not know about records for which access has been denied by the access control list server 27.

It should also be noted that the full text index 17 has no access information. Similarly, the electronic document records in storage device 25 have no security labels or information. All security information is in the access control list server 27 which relates document titles in the text documents storage device 25, their access classification, plus user identification and the access level for each user.

Example A

With reference to Fig. 2, a query server 13 has access to an access control list file which can be located anywhere, but is associated, as by a data link, with one or more web servers 23 that are indexed by the query server. The access control list has a list of all users of the system, together with a list of documents that each user is permitted to access. The access control list file may be local to the query server 13 or may be accessed remotely using a file transfer protocol (FTP). The query server uses its own filesystem file locator, 27, to access and interpret the access control list and calls up those documents in web servers 23 responsive to a search query for which the user has access. Only those documents are presented to a user.

Example B

With reference to Fig. 3, the query server 13 accesses an access control list as in the prior example, except that HTTP protocol is used instead of using the filesystem or FTP.

A particular user, Mr. Jones, ID 71234, might need access to FBI cellular communication device numbers, including pagers and telephones. A query is sent to the query server which uses HTTP protocol to access each access control list file associated with each web server whose documents are contained in the index. His access control entry would be as follows:

71234=http://federalexec.justice.gov/fbi/agents/cellular/.

He might have other entries for other classes of documents, but this class of documents will relate to FBI cellular numbers, whether pagers, telephones, or other devices. A higher level of access would be as follows:

71324=http://federalexec.justice.gov/fbi/

and a lower level of access would be as follows:

71324=http://federalexec.justice.gov/fbi/agents/cellular/pagers.html.

In the latter case, Mr. Jones would not have access to cellular telephone numbers and the web server query server would not allow access to telephone numbers. In both cases, the access control file finds the user, 71324, but in the latter case, access is denied.

Example C

With reference to Fig. 4, the query server 13 is connected to a web server 23, as before, except that the web server is running a program, for example a search program, which is triggered or controlled from the query server by a communication, which invokes a script, batch

file or executable instruction, generated by the query server and meaningful to the program on the web server where the requested document is located. Each web server whose words are indexed validates its own documents for particular users using a validation message. Such communications are known as "gateway scripts". Gateway scripts are sometimes called CGI scripts, where CGI is an acronym for Common Gateway Interface. A script may have a URL in HTTP format which controls or operates the program in the web server to execute a search query. The script can be resident in the web server and be invoked only by the URL transmitted by the query server. The search server transmits the user identification and list of candidate URLs that match the query and requests a CGI script to validate the list of URLs. The web server performs the validation and returns a list back to the search server indicating the URLs which the user is entitled to read in accord with his access level. Those documents which are beyond his access level are not reported to the user.

It is now possible to have a centralized index of documents found on multiple document servers, some or all of which may be remote. An access control list is associated with the index of documents. A search of the centralized index will report addresses, URLs, of various documents responsive to the search query. Since the access control list shows the URLs to which the user has access, only those titles to corresponding documents need to be shown to a user or fetched from a document server.

An advantage of the present invention is that changes in the access control list are immediately reflected in searches, because the list links authorized documents for each user identification code, sometimes using a hierarchical structure. In this manner, large categories of documents can be included or excluded from a search with a single file entry, such as a partial URL.